# An ontological model for digital preservation

*Panos Constantopoulos*[1,2] *and Vicky Dritsou*[1]
[1]Information Systems and Databases Laboratory, Athens University of Economics and Business
[2]Digital Curation Unit, R.C. Athena
{panosc | vdritsou}@aueb.gr

**Abstract**

The long-term preservation of digital resources requires assigning appropriate metadata. Several recommendations exist for preservation metadata reflecting the purposes and major preoccupations of the organizations that proposed each one. Digital resources are either digital surrogates of non-digital objects, or original digital objects. In either case we may assume that, if a digital resource is selected for preservation, it can be considered as a cultural object that belongs to some collection, has a certain value and must be appropriately preserved, used and documented. In other words, although the digital nature of the resource induces specific requirements in terms of preservation policies and techniques, basic commonalities can be traced with other cultural objects at the functional level. Based on this premise, the work reported here aims at harmonizing the information structures supporting digital preservation with those for documenting non-digital cultural objects. The approach taken is to develop a conceptual model for preservation metadata that complies with a standard ontology for cultural documentation, namely CIDOC CRM.

The perils faced by digital resources are related either to physical causes or to technological evolution. Correspondingly a number of preservation strategies have been developed and a preservation life cycle, independent of preservation strategy, has been identified. Furthermore, several proposals for preservation metadata have been made, five of which, appearing to be the most interesting and influential, namely the Dublin Core Metadata Element Set, the Open Archival Information System (OAIS), the Curl Exemplars Digital Archives (CEDARS) model, the Pittsburgh Project and the National Library of Australia (NLA) proposal, we analyze in order to understand their differences and to draw upon for our own model.

Our model is derived from the CIDOC CRM taking into account the above metadata sets. It is intended to capture the kinds of events occurring in the digital preservation life cycle, the kinds of objects involved and the relationships among those. Thus it could be considered as an application ontology for digital preservation derived from CIDOC CRM. It includes a minimal set of concepts appropriately inter-related and specialized. The resulting model yields a metadata set that displays significant overlap with the selected pre-existing ones. Its merit lies in the inference capability stemming from the explicit semantic structure, as well as in the integration with the domain of cultural documentation.

## 1 Introduction

Digital assets face two types of perils: physical and technical. Physical perils include various damages of storage media and catastrophic environmental incidents, e.g. fire, flooding, earthquake, etc. Preservation policies to safeguard against such perils include copying and distributing copies in different locations. These are multi-parameter policies with the details of which we are not concerned here. Technical perils include the various kinds of difficulty or inability to access and use data due to the technical evolution of hardware and software. Preservation policies against technical perils employ techniques that fall in nine main classes: migration of digital content, technology emulation, technology preservation, dedication to standards, backward compatibility, encapsulation, permanent identifiers, transformation to non-digital form and digital archaeology (National Library of Australia 1999).

The implementation of preservation policies invariably requires certain information about the digital assets, captured by preservation metadata. In this work we formulate a conceptual model for

digital preservation metadata, which (a) abstracts from certain established preservation metadata sets, (b) explicitly displays the underlying semantic relations, and (c) is compatible with CIDOC CRM (Crofts *et alii* 2004), the ISO standard ontology for cultural documentation. The latter is motivated by the fact that, in the cultural domain, preserved digital assets may be surrogates of non-digital objects and/or be cultural objects in their own right.

## 2 Digital preservation metadata sets

Several metadata sets for digital preservation have been proposed. In our work we focus on five well established ones. These are the metadata sets defined by: the Dublin Core Metadata Initiative (DCMI 2004), the Open Archival Information System (OAIS) (OCLG/RLG 2002), the Curl Exemplars Digital Archives (CEDARS) (Day 1998), the Pittsburgh Project (Archives and Museums Informatics 1996) and the National Library of Australia (National Library of Australia 1999). Here we only point at certain features of these metadata sets. Interested readers can find a detailed account in (National Library of Australia 1999) .

The Dublin Core Metadata Element Set is well documented and easy to apply, comprising only fifteen elements. Being primarily intended for supporting Web-based information access, this metadata set does not convey all the information required by preservation processes. The OAIS reference model includes four basic metadata categories: content information, preservation information, packaging information and descriptive information. All elements are hierarchically organized under these categories and their total exceeds one hundred different elements including many details. Although the model is very well documented, we believe that this big number of metadata introduces complexity in its application, compounded by the lack of specific data entry instructions. A much similar approach is proposed by the CEDARS project, also featuring high detail and difficulty of application, especially when collections of objects, rather than single files, are concerned. The significance of those detailed metadata sets lies in that they contain, though indiscriminately, all potentially useful elements. The Pittsburgh Project proposal also contains a total of more than a hundred elements, yet it makes a clear distinction between necessary and optional elements and provides clear use instructions. The National Library of Australia proposal contains twenty five basic elements, some of which are further analyzed into sub-elements. What makes this proposal remarkable is that its creators clearly state the essential sub-elements of each object, depending on its type. They analyze six object types: picture, sound, video, text, database, and executives. This is clearly an advantage, yet again no specific guidance for entering the essential data is provided.

## 3 A conceptual model for digital preservation

By comparatively studying the above metadata sets, we have identified a set of common elements we consider most significant (Dritsou 2004), see Table 1. These are elements that a conceptual model for digital preservation must contain. In addition, the model explicitly shows the semantic relations between the metadata elements, a feature we consider important for supporting the application of tools and processes. Furthermore, the model is designed as a derivative of CIDOC CRM, the now widely admitted standard cultural documentation ontology (ISO 21127). This enables interoperability with other cultural documentation data and applications.

By virtue of its derivation from CIDOC CRM, the model is *event-centric*: the concept of *activity*, which is a specialization of the *event* concept, allows for the representation of preservation processes which involve digital objects, actors, equipment and effects. Activities may be structured and form sequences. We may wish to record the actual realization of processes, a historical record, or the prescription of a process to be performed, a planning record. This can be naturally complemented by recording the deviation between plan and reality. The CIDOC CRM ontology is intended to support historical records in the above sense. Our first aim was to establish the unity of the approach to modelling digital preservation processes with the demonstrated powerful CIDOC CRM approach to modelling historical processes. Thus we gave priority to developing the comparable part of the model, i.e. the historical, with only a limited treatment of the planning part.

| |
|---|
| Title |
| Identifier |
| Subject |
| Language |
| Type |
| Format |
| Size |
| Information Carrier |
| Technical Equipment |
| Activity |
| Right |
| Actor |
| Effect |
| History |

Table 1. Basic preservation metadata elements

The model contains a minimal set of appropriately inter-related concepts, namely concepts corresponding to the metadata elements of Table 1 together with digital object, digital content and complex object as basic structural entities. The model is shown in Figure 1, where the name of each entity and property is followed by the code of the corresponding CIDOC CRM item which it is a subclass of. The elements of the model are also listed in Tables 2 and 3. Table 2 shows the entities along with their parent CIDOC CRM concept. Table 3 shows the properties along with their domain and range entities and parent CIDOC CRM property. Clearly, the majority of the model elements are defined as specializations of CIDOC CRM entities, which accounts for a high degree of compatibility between our digital preservation model and CIDOC CRM – based cultural documentation application models. The few elements with no CIDOC CRM origin are the point of departure for both differentiating and linking the (still to be developed) planning and the historical parts of the preservation model.

The central concepts of the model are represented by the entities `Digital Object` and `Activity`. The structure of `Digital Object` separates the primary content from external descriptive and administrative information. The content is held in `Digital Content`, whereas the secondary information (metadata) is captured by the attributes `Identifier` (which may be local or global), `Title`, `Subject`, `Size`, `Natural Language`, `Type` (e.g. image, text, sound, video), `Format` (a number of formats may be applicable to a certain object type, while a given object has a specific format), and `Information Carrier` (which the object is stored on). Digital objects can be composed of other objects, e.g. an html file containing both text and images. This situation is represented by `Complex Object`, a subclass of `Digital Object`.

`Activities` admit digital objects as input and produce digital objects as output. They are performed by certain `Actors` on condition these hold the appropriate `Rights`. A number of `Activity Types` are distinguished: `Creation`, `Deletion`, `Modification`, `Alteration`, `Read`, `Copy` and `Security Enforcement`, each particular activity being an instance of an activity type. `Creation`, `Modification`, `Alteration` and `Copy` produce a new object as output, while the others do not.

| Preservation Model Entity | Parent CIDOC CRM Entity |
|---|---|
| Digital Object | E73 Information Object |
| - Complex Object | E73 Information Object |
| Digital Content | E73 Information Object |
| Object Identifier | E41 Appellation |
| - Global Identifier | E41 Appellation |
| - Local Identifier | E41 Appellation |
| Size | E54 Dimension |
| Title | E35 Title |
| Subject | E1 CRM Entity |
| Natural Language | E56 Language |
| Type | E55 Type |
| Format | E29 Design or Procedure |
| Information Carrier | E84 Information Carrier |
| Technical Equipment | E71 Man-Made Stuff |
| - Software | E71 Man-Made Stuff |
| - Hardware | E71 Man-Made Stuff |
| Activity | E7 Activity |
| Activity Type | E55 Type |
| Actor | E39 Actor |
| Right | E30 Right |
| Effect | E3 Condition State |
| History | E31 Document |

Table 2. Preservation model entities and parent CIDOC CRM concepts

Preservation processes usually comprise sequences of activities. These are captured by precedence relationships represented by the property `previous`. The set of activities that have affected a given digital object, along with their sequence paths, `is documented in` a `History` entity. The particular changes that a modification, alteration or security enforcement activity has brought to a digital object are represented by `Effect`. This element is intended to explicitly record object changes, whether restoration to the previous condition is possible or not. Furthermore the use of `Effect` is independent of the possible tracking of versions, the latter actually remaining out of the scope of the preservation model. Finally, all activities require `Technical Equipment` in order to be carried out, `Hardware` and `Software`. Given the `Format` of a digital object, the `Software` it `is supported by` needs to be specified.

The preservation conceptual model can be considered as an application ontology derived from CIDOC CRM. The majority of concepts in this model have been defined as specializations of CIDOC CRM concepts. In addition certain independent extensions have been necessary. The `has effect` property of `Activity` with range `Effect` is one presented above. Another is the `is formatted in` whereby `Format` is related with `Digital Object`. On the other hand `uses format` associates `Type` with `Format` to indicate the possible formats a digital object of a given type may be in.

| Property Name | Domain | Range | Parent CIDOC CRM Property |
|---|---|---|---|
| contains | Digital Object | Digital Content | E73 Information Object. P106 is composed of (forms part of): E73 Information Object |
| is identified by | Digital Object | Object Identifier | E1 CRM Entity. P1 is identified by (identifies): E41 Appellation |
| has size | Digital Object | Object Identifier | E70 Stuff. P43 has dimension (is dimension of): E54 Dimension |
| has title | Digital Object | Title | E71 Man-Made Stuff. P102 has title (is title of): E35 Title |
| has subject | Digital Object | Subject | E73 Information Object. P129 is about (is subject of): E1 CRM Entity |
| has language | Digital Object | Natural Language | E33 Linguistic Object. P72 has language (is language of): E56 Language |
| has type | Digital Object | Type | E70 Stuff. P101 had as general use (was use of): E55 Type |
| is saved to | Digital Object | Information Carrier | E24 Physical Man-Made Stuff. P128 carries (is carried by): E73 Information Object |
| is formatted in | Digital Object | Format | --- |
| is supported by | Format | Software | --- |
| carries out | Actor | Activity | E7 Activity. P14 carried out by (performed): E39 Actor |
| is subject to | Digital Object | Right | E72 Legal Object. P104 is subject to (applies to): E30 Right |
| held by | Right | Actor | E39 Actor. P75 possesses (is possessed by): E30 Right |
| to perform | Right | Activity | --- |
| takes as input | Activity | Digital Object | E7 Activity. P16 used specific object (was used for): E70 Stuff |
| gives as output | Activity | Digital Object | E81 Transformation. P123 resulted in (resulted from): E77 Persistent Item |
| hat type | Activity | Activity Type | E7 Activity. P21 had general purpose (was purpose of): E55 Type |
| requires | Activity | Technical Equipment | E7 Activity. P16 used specific object (was used for): E70 Stuff |
| has effect | Activity | Effect | --- |
| previous | Activity | Activity | E7 Activity. P134 continued (was continued by): E7 Activity |
| documented in | previous | History | E31 Document. P70 documents (is documented in): E1 CRM Entity |

Table 3. Preservation model properties and parent CIDOC CRM properties

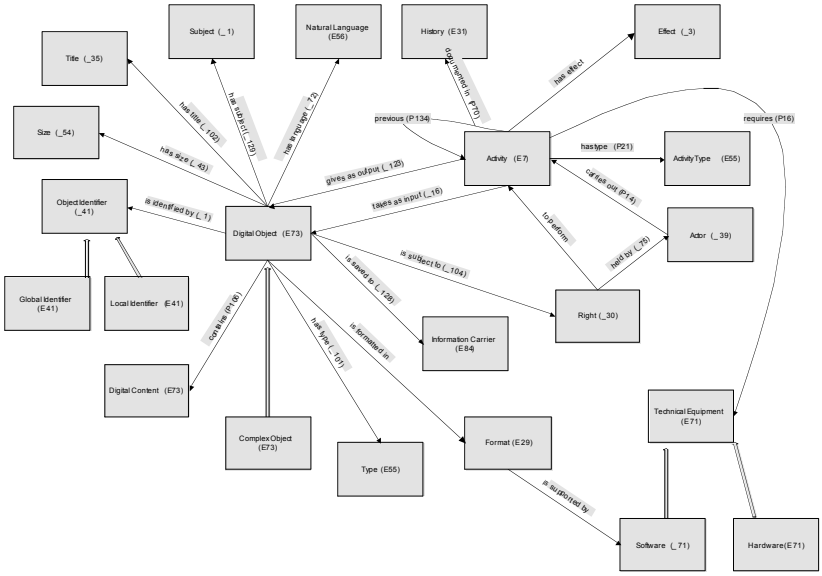


Figure 1: Conceptual model for digital preservation

# 4 Conclusion

We have presented a conceptual model for representing digital preservation processes, which draws elements from established preservation metadata sets and complies with the ISO standard cultural documentation ontology CIDOC CRM. Thus it could be considered as an application ontology for digital preservation derived from CIDOC CRM. It includes a minimal set of concepts, appropriately inter-related, and yields a metadata set that displays significant overlap with the selected pre-existing ones. Its merit lies in the inference capability stemming from the explicit semantic structure, as well as in the integration with the domain of cultural documentation.
Around the basic concept of activity, this model allows for representing both historical processes and planning processes. The historical part is the one already developed and connected with CIDOC CRM. In what concerns the planning part, now a potentiality, we contend it would be interesting to examine the planning processes in further detail and explore the possible differences in modelling requirements that may arise from the fact that these are actually decision and production processes explicitly documented for operational purposes, whereas the historical processes accounted for in the ex post documentation of acquired cultural objects are approached interpretatively.

## References

Archives and Museum Informatics 1996. “Metadata Specifications Derived from the Functional Requirements: A Reference Model for Business Acceptable Communications”. www.archimuse.com/papers/nhprc/meta96.html

Bekiari, C., Constantopoulos, P., Doerr, M. (eds.) 2006. Cultural Documentation and Interoperability Guide (in Greek). www.ics.forth.gr/CULTUREstandards

Cornell University. “Metadata Types”. www.library.cornell.edu/preservation/tutorial/matadata/table5-1.html

Coyle, K. 2004. “Metadata: Data with a Purpose”. California Library Association Meeting. www.kcoyle.net/meta_purpose.html

Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M. 2004. CIDOC CRM. http://cidoc.ics.forth.gr/official_release_cidoc.html

Day, M. 1998. Metadata for Preservation. CEDARS Project Document AIW01. www.ukoln.ac.uk/metadata/cedars/AIW01.htm

DCMI 2004. Metadata Terms. dublincore.org/documents/2004/09/20/dcmi-terms/

Digital Preservation Coalition. Digital Preservation Coalition Handbook. www.dpconline.org/graphics/handbook

Dritsou, V. 2004. Digital Content Preservation Metadata. M.Sc. in Information Systems. Athens University of Economics and Business.

Hodge, M.G. 2000. "Best Practices for Digital Archiving: An Information Life Cycle Approach". In D-Lib Vol. 6 No. 1. www.dlib.org

National Library of Australia 1999. Preservation Metadata for Digital Collections. www.nla.gov.au/preserve/pmeta.html

OCLC/RLG Working Group on Preservation Metadata 2002. A metadata framework to support the preservation of digital objects. www.oclc.org/research/projects/pmwg/pm_framework.pdf